AFRL-IF-RS-TR-2006-79
Final Technical Report
March 2006

PHYSICO-CHEMICAL PROKARYOTE MODELS: STAND-ALONE MODULES AND KARYOTE INTEGRATION

Indiana University

Sponsored by
Defense Advanced Research Projects Agency
DARPA Order No. P127

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2006-79 has been reviewed and is approved for publication.

APPROVED: /s/

ROBERT L. KAMINSKI
Project Engineer

FOR THE DIRECTOR: /s/

WARREN H. DEBANY, Technical Advisor
Information Grid Division
Information Directorate

DESTRUCTION NOTICE - For classified documents, follow the procedures in DOD 5200.22M. Industrial Security Manual or DOD 5200.1-R, Information Security Program Regulation. For unclassified limited documents, destroy by any method that will prevent disclosure of contents or reconstruction of the document.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	MARCH 2006	Final Sep 02 – Sep 05

4. TITLE AND SUBTITLE
PHYSICO-CHEMICAL PROKARYOTE MODELS:
STAND-ALONE MODULES AND KARYOTE INTEGRATION

5. FUNDING NUMBERS
C - F30602-02-2-0001
PE - 61101E
PR - BIOC
TA - P1
WU - 27

6. AUTHOR(S)
P. Ortoleva

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Indiana University
P. O. Box 1847
Bloomington Indiana 47402-1847

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 North Fairfax Drive
Arlington Virginia 22203-1714

AFRL/IFGA
525 Brooks Road
Rome New York 13441-4514

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2006-79

11. SUPPLEMENTARY NOTES
AFRL Project Engineer: Robert L. Kaminski/IFGA/(315) 330-1867/ Robert.Kaminski@rl.af.mil

12a. DISTRIBUTION / AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***
At the Center for Cell and Virus Theory (CCVT) several types of systems biology modules were developed for the Bio-SPICE project. The cell modeling modules account for metabolic, proteomic and genomic kinetics and their spatially localized, multiple scale character. Modules for model development, calibration and multiplex (e.g. genome-wide expression) data interpretation are also provided. Models are made available via three complementary mechanisms: 1) the Bio-SPICE system; 2) open source stand-alone code; and 3) a website (sysbio.indiana.edu) run by CCVT. The latter can also be run through the Bio-SPICE Dashboard. CCVT software has been demonstrated in a variety of contexts including transcriptional regulatory networks in prokaryotes and B cells, self-organized division mechanisms in E. coli, glycolysis in yeast and T. brucei, and the transcriptional response of human cells subjected to toxins. As CCVT also has an educational function, a number of graduate and undergraduate students have been trained as the next generation of systems biologists.

14. SUBJECT TERMS
Bio-SPICE, gene expression, cell modeling, data/model integration, kinetics, diffusion, transcription factor, transcriptional regulatory network

15. NUMBER OF PAGES
35

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Table of Contents

List of Figures

List of Tables

1.0 Summary

Systems biology modules were developed for Bio-SPICE. The modules were of two types – cell models and data/model integrators. Selected modules were made available through Bio-SPICE as components of the Dashboard while others were designed to be stand-alone or available as a web service.

Cell models are Karyote and CellX. The former is an ordinary differential equation model of a compartmented cell which can simulate both eukaryotic and prokaryotic cells. The reactions allowed in each compartment and the membrane transport processes accounted for are general in character. A companion website (*sysbio.indiana.edu*) was set up that allows users to create files for Karyote input. These include single and multiple cell (suspension and tissue) models. Karyote was demonstrated using yeast and the parasite *T. brucei*, the causative agent in sleeping sickness.

CellX is a partial differential equation-based cell model. In its Bio-SPICE implementation it is designed for prokaryotic cell simulation. Reaction-transport equations are solved in the cell interior (3-D), on or within the cell membrane (2-D), and a boundary continuity equation accounts for processes of exchange between the interior and the membrane. CellX was demonstrated for the self-organized plane of division in *E. coli* via Min protein reaction-transport processes.

Data/model integration modules were also developed. These modules allow a user to directly extract cell modeling information from multiplex data (e.g. cDNA microarray, NMR and proteomics). The microarray-based modules were made part of Bio-SPICE. One of these modules, FTF, is designed to extract transcriptional regulatory information from cDNA microarray data in time series or steady states for cells in various extracellular conditions. The KAGAN module uses cDNA microarray data to refine a transcriptional regulatory network and calibrate associated rate and binding constants. Our transcriptional regulatory network construction modules (FTF and KAGAN) are built on the estimation of transcription factor profiles. Most other methods of microarray data analysis are based on the assumption that protein profiles are in step with profiles of the encoding RNA – an assumption that has been shown experimentally to be untrue in a number of cases where detailed proteomics and RNA expression data were both available.

FTF is highly CPU efficient so that many networks can be tested to arrive at one which is most consistent with the available microarray data. We demonstrated that FTF is ideally nested in an overall workflow aimed at transcriptional regulatory network construction. In this way FTF, combined with promoter analysis, gene ontology and phylogenic similarity can be used to greatly increase the number of transcriptional factor/gene regulatory interactions discovered. FTF with gene ontology was used with a dataset of 336 microarrays on B cells to create a very large network for these cells (posted at *sysbio.indiana.edu*).

KAGAN, the second microarray-based transcriptional regulatory network construction module developed and installed at Bio-SPICE, can refine an input network and calibrate the transcription and RNA degradation rate constants, as well as transcription factor/gene binding constants.

As FTF is optimized for network construction, and KAGAN is designed to take a network and refine and calibrate the associated biochemical kinetic parameters, they are ideally suited for a two step network inference workflow. This has been implemented at *sysbio.indiana.edu*.

For FTF, KAGAN, and our bioinformatics modules, it is necessary to have a preliminary network/training set. To serve users we have created the GeneDat database containing over 13,000 transcription factor/gene regulatory interactions for mammalian (mostly human) cells. It also contains what we believe to be state-of-the-art network information for *E. coli* and *B. subtilis*. With each transcription factor/gene regulatory interaction a variety of annotations are provided.

2.0 Introduction

The modeling systems developed in this Bio-SPICE project have distinct levels of physics and chemistry according to the phenomena they address. In this report the Bio-SPICE modules developed at CCVT are described in some detail. There modules are as follows.

The Karyote system solves ordinary differential equations: a cell is divided into compartments within each of which user-specified reactions take place and between which molecules are exchanged by active and passive processes. A detailed tutorial with 15 instructive mechanistic models is provided as are datasets for yeast glycolysis and *T. brucei* (the causative agent of sleeping sickness) oxidative and anoxic glycolysis. Karyote has special features such as the ability to construct models involving suspensions of various types of cells (e.g. blood) and the unification of two cell models into one, more comprehensive one. Karyote is SBML compatible (i.e. SBML files can be created for use by other simulators or received from others as input files).

FTF and KAGAN are unique microarray data analysis modules for the construction of gene regulatory networks. FTF focuses on network structure (i.e. it delineates the transcription factors regulating each gene). KAGAN focuses on calibrating rate coefficients for transcription and RNA degradation, transcription factor/gene binding constants, and network structure refinement.

GeneDat is a database of over 13,000 experimentally-verified up/down transcription factor/gene regulatory interactions (mostly human) annotated with species, cell line, and data source. Software associated with GeneDat launches queries that automatically provide a preliminary regulatory network for FTF, KAGAN, or other network analysis/improvement modules.

CellX has most of the features of Karyote. In addition, 3-D concentration distributions within the cell and 2-D distributions along the cell inner surface are computed using finite element methods.

3.0 Karyote Cell Analyzer

Overview

The Karyote Cell Analyzer (KCA) is based on the principles of chemistry and physics as formulated for single and multiple cell systems. It is designed for fundamental studies into the workings of a cell and for applications of this understanding in drug and vaccine design, treatment optimization, refined diagnoses of complex diseases like cancer, environmental analysis, and biotechnical process engineering. The input to KCA is the network of biochemical processes and rate/equilibrium data, intracellular architecture, membrane transport properties, initial cell state and conditions in the extracellular medium. The output of KCA is the concentration timecourse of all chemical species in all compartments (and all cells for a multicellular simulation).

Karyote Cell Analyzer Functionality

KCA simulates compartmentalized cellular reaction-transport processes. Ordinary differential equations are solved in each compartment and active and passive molecular transfer between compartments is accounted for. All processes are fully coupled through the dependence of rate laws on composition. Solution techniques include multiple timescale analysis and a stiff solver package with method switching for efficiency. Reactions designated as fast are maintained close to equilibrium or steady state as coupled processes may indicate. Mass conservation errors made by assuming rate laws in the form of polynomial ratios are avoided. Both rate and equilibrium constants must be supplied for finite rate processes while only equilibrium constants are to be provided for fast, equilibrated ones; when subsets of fast reactions are in steady state balance (with a nonzero net overall rate) both forward and reverse rate constants must be provided.

Many physico-chemical processes underlying cell behavior are accounted for in KCA. In these notes we familiarize the user with them by illustrating how to run KCA via examples as described in *Explaining KCA Though Simple Cell Models*, below. In KCA one may consider a cell to be a single compartment reactor or may divide it into compartments within each of which specialized reactions take place. One may build very extensive reaction networks, including equilibrated reactions or cycles of reactions in steady-state balance that have great complexity. One may vary parameters in the membrane flux laws or create transcompartmental reactions that simultaneously involve chemical species on both sides of a membrane (e.g. catalysis or ion

pumps). As KCA computations are hierarchical one may construct systems with compartments within compartments and thereby model multi-cellular systems (e.g. cell suspensions, tissues or embryos).

Quantity	Units
Time	seconds
Volume	liters
Area	$(\text{decimeters})^2$
Concentration	millimole/liter
equilibrium constant for *m*-th order forward, *n*-th order reverse action	$(\text{millimole/liter})^{n-m-1}\text{sec}^{-1}$
rate coefficient for *n*-th order reverse action	(liter/millimole)n-1
K_M membrane permeability factor	millimoles/liter
V_M maximum permeability	decimeter (sec)-1
ϕ membrane asymmetry parameter	none

Table 1. Units for input or output variables in KCA.

The nature of a KCA cell model is suggested in Fig. 1. The system is divided into compartments (i.e. cytosol and organelles). These subsystems are separated by membranes across which molecules can be exchanged. Molecules can only exchange between compartments that the user specifies to have a common membrane of given surface area; thus surface areas $A^{\alpha\alpha'}$ between compartments α and α' are used to define system configuration, and similarly for the volume V^{α} of compartment α. This allows the user to define complex intra-cellular architectures or multi-cellular systems (i.e. consisting of compartments within compartments in a hierarchical fashion). Many of the general chemical kinetics concepts used in KCA are reviewed in the following sections and in a mini-course on Chemical Kinetics available at *sysbio.indiana.edu*.

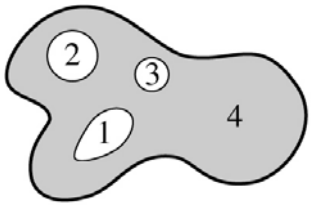


Figure 1 – Cell divided into compartments

Figure 1: In Karyote the cell is divided into compartments labeled $\alpha = 1,2,\cdots$ separated by membranes. For each compartment ordinary differential reaction-transport equations are used to simulate the evolution of descriptive variables (e.g. concentrations and electrical potentials).

Mathematical Formulation

Publications that contain more technical details about the KCA include Ortoleva et al. (2003); Weitzke and Ortoleva (2003); Sayyed-Ahmad et al. (2003); and Navid and Ortoleva (2004). These and other papers are available through our website (*sysbio.indiana.edu*).

In KCA the cell is divided into N_C compartments labeled $\alpha = 1,2,\cdots N_C$. In compartment α, each molecular species $i = 1,2,\cdots N$ is described by its concentration c_i^α. Conservation of mass implies

$$V^\alpha \frac{dc_i^\alpha}{dt} = \sum_{\alpha' \neq \alpha}^{N_c} A^{\alpha\alpha'} J_i^{\alpha\alpha'} + V^\alpha \sum_{l=1}^{N_r} \nu_{il}^\alpha \, W_l^\alpha \,. \tag{1}$$

$A^{\alpha\alpha'}$ = boundary surface area separating compartments α and α'

$J_i^{\alpha\alpha'}$ = net flux of species *i* from α' to $\alpha \left(= -J_i^{\alpha\alpha'}\right)$

N, N_c, N_r = number of chemical species, compartments and reactions, respectively

V^α = volume of compartment α

W_k^α = rate of reaction *k* in compartment α

ν_{ik}^α = stoichiometric coefficient for species *i* in reaction *k* in compartment α.

In KCA the last term is divided into fast and slow contributions to take advantage of multiple scale methods. Let ε be a small parameter. Then the net reaction rate (the last term in (1) divided by V^α), denoted R_i^α, is written $R_i^\alpha = \sum_{l=1}^{N_s} \nu_{il}^{\alpha s} W_l^{\alpha s} + \frac{1}{\varepsilon} \sum_{l=1}^{N_f} \nu_{il}^{\alpha f} W_l^{\alpha f}$.

As $\varepsilon \to 0$ the system is driven close to equilibrium (so that $W_l^{\alpha f} = 0$) or linear combinations of the $W_l^{\alpha f}$ vanish (to express steady-state cycles). Furthermore minority species (e.g. enzymes) impart another element of stiffness to the cell simulation problem. Thus in KCA species are divided into majority and minority categories. The latter have concentrations that scale with ε in our formalism. The resulting multi-scale analysis avoids difficulties in numerical simulations, especially when minority species participate in fast reactions. This minority/majority separation

allows for greater computational efficiencies. As the use of this option places more burden on the user (i.e. discriminating minority versus majority species), we do not include it in this Bio-SPICE release.

Rates of reaction are of the mass-action form:

$$Rate = k\left[Q\prod_{\nu_i>0} c_i^{\nu_i} - \prod_{\nu_i<0} c_i^{-\nu_i} \right] \tag{2}$$

where k is the reverse rate coefficient, Q is the equilibrium constant (i.e. kQ is the forward rate coefficient) and ν_i is the stoichiometric coefficient (i.e. $\nu_i > 0$ for products and < 0 for reactants. The passive flux law is in the form

$$J^{passive} = h\left[c' - c\right] \tag{3}$$

$$h = \frac{K_M^2 V_M}{\left[K_M^2 + K_M\left(c + c'\right) + \phi c c'\right]} \tag{4}$$

for a given membrane and species; c and c' are concentrations on either side of the membrane (see Table 1 for other variables).

Intramembrane enzymatic and active transport processes are accounted for as transcompartmental reactions. For example,

$$2A\left(cytosol\right) + B\left(mitochrondrion\right) \rightleftharpoons C\left(cytosol\right) + 3D\left(mitochrondrion\right)$$

Again the rate law is assumed to be of the mass action form. The dependence of the rate of such processes on the area of the membrane is accounted for, i.e. the transcompartmental reactions and passive flux law contributions are added in computing $J_i^{\alpha\alpha'}$.

The user can choose between explicit integration, Runge-Kutta integration (both using multiple timescale separation techniques discussed in Ortoleva (1992) and Weitzke and Ortoleva (2003)), and implicit integration. We use the double precision VODE (*www.llnl.gov/CASC/download/download_home.html*) as the implicit integrator.

The Web-based Karyote Cell Modeling System

The Karyote Cell Modeling System (KCMS) performs a multiplicity of functions in support of DARPA Bio-SPICE activities by familiarizing beginning and advanced researchers with

concepts in cell modeling. As new computational biology modules are developed and tested at CCVT they are in the web-based system at *sysbio.indiana.edu*.

Essential features of the web-based KCMS system include:

- Cell Assembler (to build a cell model and create an SBML KCA input file);
- Cell Unification (to integrate existing cell models, pathways or subsystems into a more comprehensive model and create an SBML file);
- Multi-Cellular System (to build a suspension or other multi-cellular configuration from single cell models and create an SBML file);
- Information Theory (to calibrate a model using NMR, spectral, microarray, electrical potentiometry and other datasets individually or simultaneously in static or time series – only microarray data analysis is installed at this writing); and
- Run the Cell Model (to derive information about the response of a cell to changes in extra-cellular conditions and other stimuli or interaction with other cells).These functions are continuously being improved.

4.0 Transcriptional Regulatory Network Construction

Overview

Two modules for constructing transcriptional regulatory networks using cDNA microarray data were developed for Bio-SPICE. In this section they are described and it is shown how they can be used in a wider strategy that introduces several bioinformatics modules in order to close the gap between the great complexity of a transcriptional regulatory network and the information needed to do so.

A number of techniques have been proposed to infer transcriptional regulatory networks (TRNs) using cDNA microarray-monitored expression profiles. Among them are principal component analysis (Holter et al. 2000, 2001) and independent component analysis (Liebermeister 2002). Network component analysis (NCA) differs from other techniques in that the structure of the gene regulatory network is assumed to be known (Liao et al. 2003). Therefore, NCA's use is limited to cases in which the network is fairly well known and has strong structural limitations. In reality, only an incomplete and possibly biased TRN is available for any cell due to the experimental conditions imposed. Gardner et al. (2003) proposed a methodology to construct the gene-gene control network structure of small networks using microarray data, limiting the number of interactions per gene. We tested a similar approach for large networks and showed that even when there are just a few interactions per gene, there can be thousands of networks that are consistent with a given microarray dataset to within essentially the same accuracy. Kyoda et al. (2000) developed a methodology that employs mutation experiments to arrive at the TRN. However, it is questionable whether their approach can be applied to large TRNs. Liang et al. (1998) presented a methodology for Boolean networks and applied it to a small 50 gene system with at most 3 interactions per gene. Boolean networks are an oversimplification of gene expression as they use a binary approximation fully on or off (Huang 1999). Cluster analysis is based on statistical techniques wherein correlations are sought between the responses of genes (e.g. Azuaje 2002; Bolshakova and Azuaje 2003). However the coordination can be extremely complex and circuitous, i.e. genes may be involved in a multi-branch feedback loop involving several transcription factors (TFs) made, or activated/deactivated by their resulting translated proteins. These time-delayed, complex relationships are revealed by our method as it discovers and quantifies many of these feedback relationships. Although cluster

analysis might suggest groups of genes that have related functionalities, it is not an accurate methodology for suggesting TF/gene regulatory interactions. D'Haeseleer et al. (2000) applied clustering based on correlation of microarray data. To assess the feasibility of inferring networks using expression data only, we used two independent gene expression datasets and a TRN for *E.coli* (*http://ecocyc.com*). Fig. 2 shows the probability of correlation between two random genes and that for known TF/gene interactions. The similarity of these distributions demonstrates that a successful reconstruction of the network using expression data alone does not seem likely. Mutual information (Basso et al. 2005) seems to have similar limitations.

If TRN construction from microarray data is unfeasible because of the insufficient information in this data, then the solution is to use as much additional information as possible to rule out spurious networks. Segal et al. (2003) assumed that genes in the same pathway are similarly regulated and their protein products often interact. This led them to the use of protein-protein interaction information in their predictions. Brazma et al. (1998) studied the similarities of the upstream regions of genes that have a similar expression profile. A similar study was presented by Haverty et al. (2004) who used statistical methods for identifying overabundant TF binding motifs (from TRANSFAC and JASPER) and microarray data to infer the TRN. The methodology we have developed is the only one that computes TF activity profiles, correlates them with microarray monitored RNA profiles, and integrates the results with promoter, gene ontology, and phylogenic analyses as follows.

Network inference using a similarity measure assumes that the activity of a TF is represented by the expression of its encoding gene. Failure to observe such a high correlation for *E. coli* (Fig. 2) shows that this assumption does not hold. Therefore, in order to use expression data to construct a TRN, we estimate TF activities independent of the expression profile of the encoding gene. This is a major shift in strategy. Our approach not only suggests highly probable TF/gene interactions, but also TF activities which can be used to establish the sense (up versus down) of the regulation and to explore post-translational reactions that create or modify TFs.

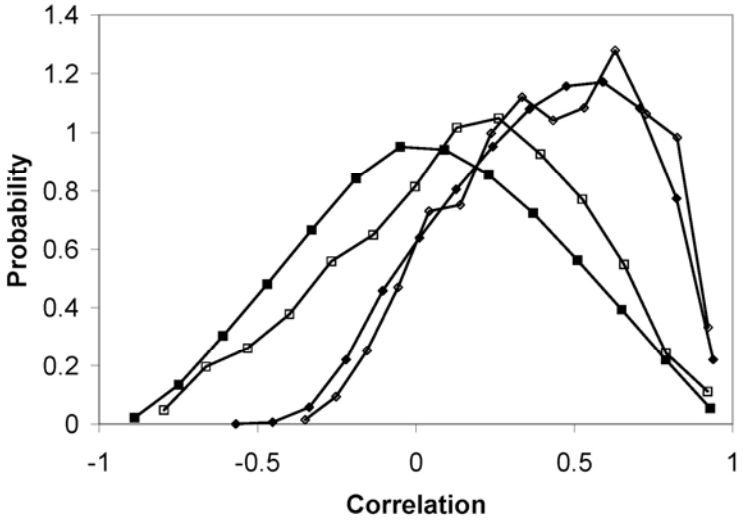


Figure 2: Probability distribution for correlation (Pearson) between a random pair and known F/generegulatory interaction for E.coli.

Square markers refer to the dataset obtained from the U. of Oklahoma E.coli database (89 datasets; http://chase.ou.edu/macro/). Diamond markers refer to the datasets obtained from the NIH omnibus service (GSE7, GSE8, GSE9; 65 datasets). The solid and hollow markers show the probability distribution for correlation between a random gene pair and known gene/TF regulatory interaction, respectively. As these probability distributions are indistinguishable, it does not seem feasible to construct the TRN using expression data alone. We also calculated probability distributions for mutual information which yielded similar findings.

FTF: A Statistical Approach to Estimate TF Activity Profiles

In designing our microarray-based TRN discovery approach, we addressed the following challenges:

- omnipresent noise/uncertainty in the data;
- vastness of the TRN;
- many regulatory mechanisms (e.g. from TF/gene binding to phosphorylation and histone interactions); and
- sparseness of the data (relative to the vastness of a TRN) imposed by the cost of microarray data acquisition.

Thus, we have developed FTF (Fast Transcription Factor) for network construction via TF activity estimation, statistical arguments, and a preliminary TRN. FTF is based on the following notions:

- gene expression data is usually error-prone and thus some consensus method is needed whereby results from a variety of genes are synthesized to derive regulatory information on a given gene;
- a method based on TFs has the advantage that microarray noise, and error in a user-supplied TRN, can be overcome by statistics ─ i.e. the regulation of many genes by a given TF;
- due to data uncertainty sparsity, there is not usually sufficient information to simultaneously obtain the structure of the TRN and the associated transcription and RNA degradation rate coefficients (as needed in steady-state or time-dependent chemical kinetic methods);
- network discovery requires many automated trials of possible networks to identify those that are most consistent with the data, so the algorithm must be extremely efficient; and
- thus the objective of FTF is to discover the structure of the TRN by taking advantage of the statistical robustness allowed by a TF-based statistical analysis.

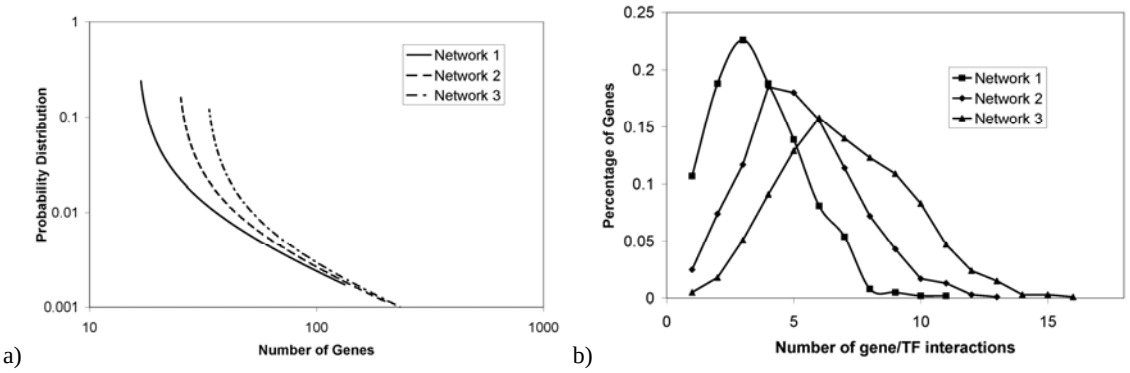
The essential equation on which FTF is based was arrived at empirically after extensive numerical experimentation with synthetic data for which we know the TRN, TF activities, and the statistics of noise added to the expression data:

$$T_n^r - T_n^s = \sum_{i=1}^{N_g} H\left(m_i^r - m_i^s\right) b_{in} \Psi_{in} , \tag{5}$$

where T_n^r = activity of TF *n* at condition or time r, m_i^r = microarray response for gene i at condition r, b_{in} = TRN ($b_{in} = +1/-1$ for gene i up/down regulated by TF n, $b_{in} = 0$ for no regulation), $H(x) = \pm 1$ for $x >$ or < 0, $= 0$ for $x = 0$, and $\Psi_{in} = 2^{L_i} / (M_n 2^{L_i - 1})$ for L_i = number of TFs controlling gene i, and M_n = number of genes TF *n* regulates. If there are N_{cDNA} time points or conditions, then one can write $N_{cDNA} \times (N_{cDNA} + 1)/2$ equations for the N_{cDNA} activities T_n^r $r = 1, 2, \cdots N_{cDNA}$, for each of the N_{TF} TFs. Therefore TF activities are obtained from the solution of (5) via a least squares fit.

Our synthetic examples with large TRNs show that, despite the simplicity of this approach, the constructed TF activities are reliable. For example, for a TRN that has the properties shown in Fig. 3, even when we eliminate 50% of the TRN to create a "preliminary TRN", 90% of the

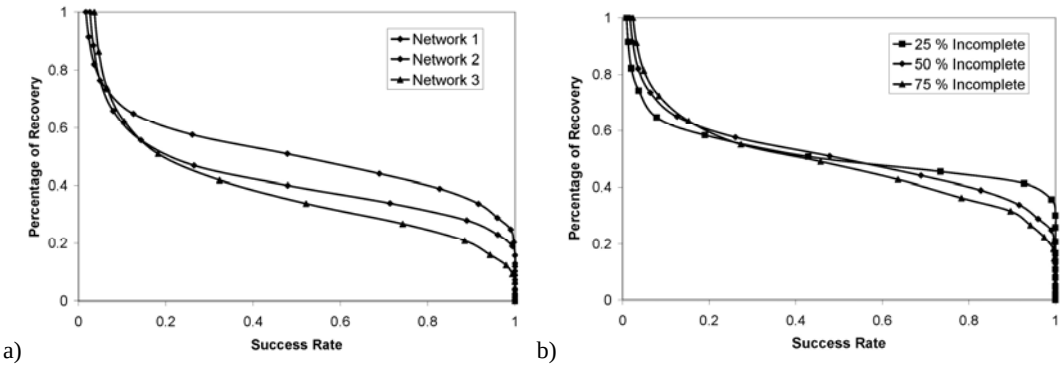
constructed TF activities have a correlation coefficient of at least 0.70 with the actual TF activities used to generate the synthetic expression data (with 20 or more microarray experiments).



a) b)

Figure 3: Properties of TRNs used in the synthetic examples.

Networks that consist of 1000 genes and 100 TFs are generated using the probability distribution for the number of genes regulated by a TF shown in (a). The corresponding probability distribution for the number of regulators per gene is shown in (b). The average number of regulators per gene is 3.62, 5.22, and 7.02 for Networks 1, 2 and 3, respectively. Equal likelihood is chosen for up/down regulation.



a) b)

Figure 4: Effect of TRN properties.

We used Networks 1, 2 and 3 of Fig. 3 to generate 100 synthetic expression datasets, and eliminated 50% of the TF/gene interactions in the TRN. Shown is the percentage of the deleted network recovered as a function of success rate. As the number of TF/gene interactions increases, percentage of the network that can be recovered decreases. b) Same as a) except we used Network 1 and eliminated 25%, 50%, and 75% of the network. As one would intuitively expect, higher percentage of the deleted network is recoverable when a more complete network is known.

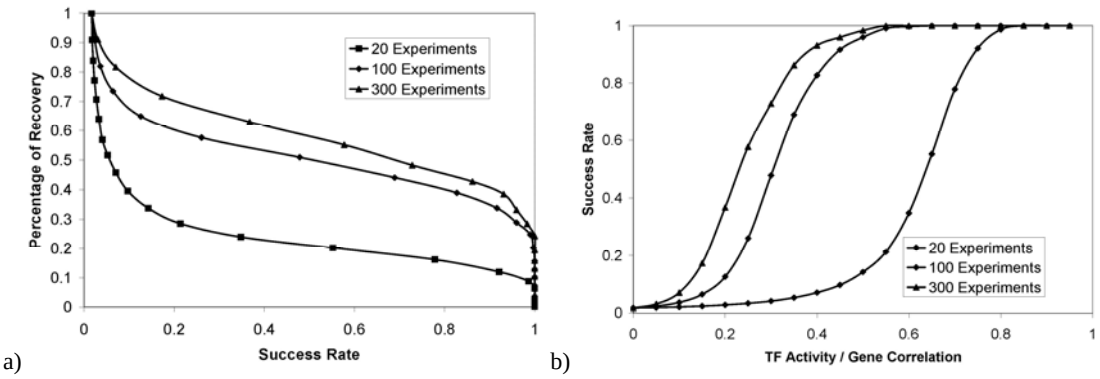


Figure 5: Reconstruction of TRNs.

We have used the Network 1 of Fig. 3 and generated synthetic expression data. Then, we eliminated 50% of the network (randomly), and used FTF to reconstruct the deleted network. Fig. a) shows the percentage of the deleted network recovered as a function of success rate, a measure of the likelihood that an interaction is correct, as estimated from the training set (known interactions). As the number of microarray experiments increases, a higher percentage of the network can be reconstructed. However, full reconstruction requires too many experiments. Fig. b) shows success rate as a function of the absolute value of the linear correlation between the constructed TF activities and gene expression data.

The essence of this approach is to estimate TF activities from a preliminary TRN (training set) and expression data. Once approximate TF activities are constructed, we calculate their correlation with the expression profiles of the genes they might regulate, and rank plausible TF/gene interactions. Results from synthetic examples using a network of 1000 genes and 100 TFs are encouraging (Figs. 4-5).

Gene Ontology

We use the biological process ontology developed by the Gene Ontology (GO) consortium (*www.geneontology.org*) and hypothesize that a gene pair is more likely to be regulated in the same manner as the similarity between their GO descriptions increases. As a gene product might be assigned multiple GO terms, we use the maximum similarity between all possible combinations. Use of GO similarity has already been shown to provide information about functional modules in *E.coli* (Wu et al. 2005). We have extended this methodology to construct to construct TRNs. Details on integration of GO into a TRN construction strategy are given below.

Phylogenic Similarity Analysis

In the phylogenetic profile algorithm, we first seek orthologous genes (in a set of *N* sequenced and annotated bacterial genomes) for each gene in the bacterium of interest. For each gene, we

construct a vector of length *N* whose *i*-th element is assigned *0* (no orthologous gene is found in bacterium *i*) or *n* (orthologous gene is found in bacterium *i* and its order in the genome is *n*). The hypothesis is that if two vectors (for a gene pair) show a high level of similarity, this gene pair is likely to be similarly regulated. In our implementation we use 230 genomes from *ftp://ftp.ncbi.nih.gov/genomes/Bacteria/*. As operons are an important feature of bacterial genomes, this approach is very likely to provide functional relationships as already shown for *E.coli* (Wu et al. 2005). The vectors noted above can be used to develop various measures of similarity that yield a probability for the accuracy of any suggested TF/gene regulatory interaction discovered. TRN enhancement for phylogenic similarity and GO, any regulatory information from a given gene is taken to be allocable to another which has a high phylogenetic similarity score.

Multiple TRN Construction Methodology

Each of the above individual method provides a score for each suggested TF/gene interaction. The statistical significance of the score is assessed by the ratio of the probability of that score in the training set to that in the random set. To meet the objective of genome-wide TRN discovery we seek an approach that integrates sufficient information to delineate the many TF/gene interactions and to eliminate spurious ones. In an attempt to develop an objective integration of the three methods (FTF, phylogenic similarity, GO) for a united TRN discovery workflow, we hypothesize that the sum of the logs of the Bayesian-like ratios for a given TF/gene interaction provides a reliable success measure. Application of the approach to *E.coli* is shown in Fig. 6. The results are posted at *sysbio.indiana.edu*. The microarray dataset was gathered from NIH omnibus service (GSE7, GSE8, GSE9; 65 datasets).

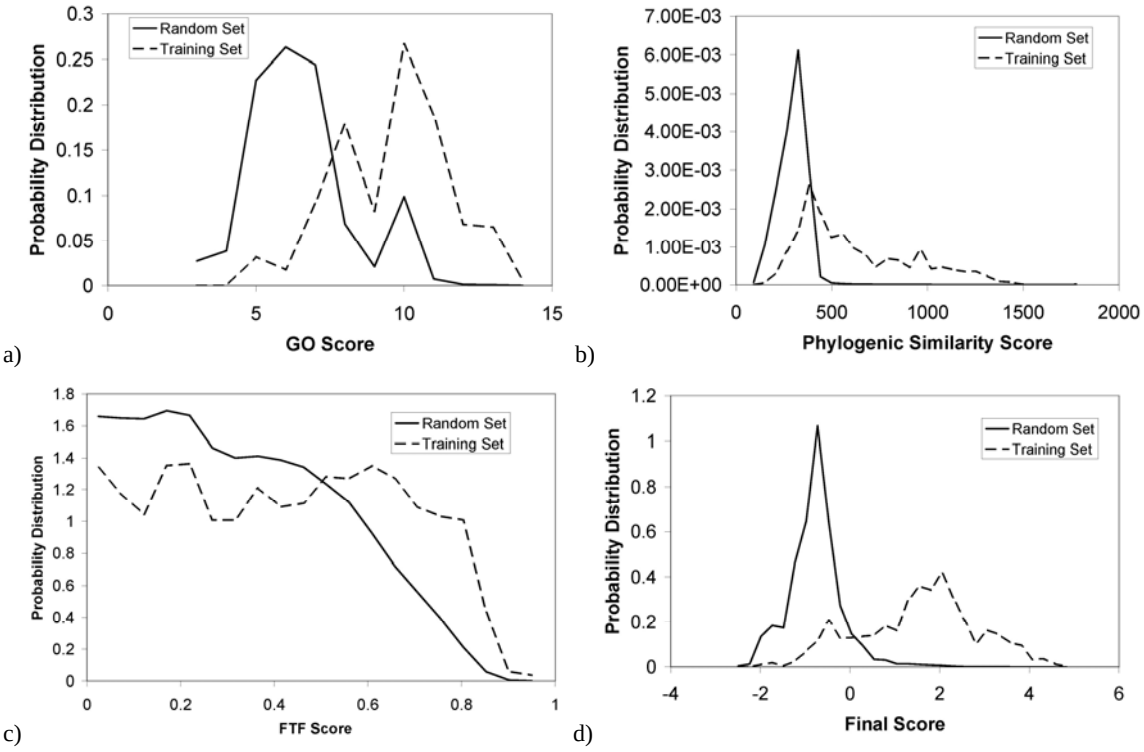


Figure 6: Probability distribution as a function of a) GO, b) phylogenic similarity, c) FTF scores.

In comparison with Fig. 2 (correlation based on gene-gene correlations), the probability distributions for the training and random sets can be easily distinguished. d) the probability distributions for the logarithm of the multiplication of Bayesian ratios.

We have applied this methodology using the 336 microarrays on B cells (GEO: GSE 2350, submitted by K. Basso) using a preliminary TRN constructed from our GeneDat database, FTF and GO similarity. Over 15,000 predicted TF/gene interactions were discovered and are posted at http://systemsbiology.indiana.edu. The final comparison of the probability distributions for the training and random sets are shown in Fig. 7. The B Cell microarray dataset was also used by Basso et al. (2005) to make predictions for the MYC TF. Our prediction set spans 489 TFs for which we had a preliminary TRN in the GeneDat database.

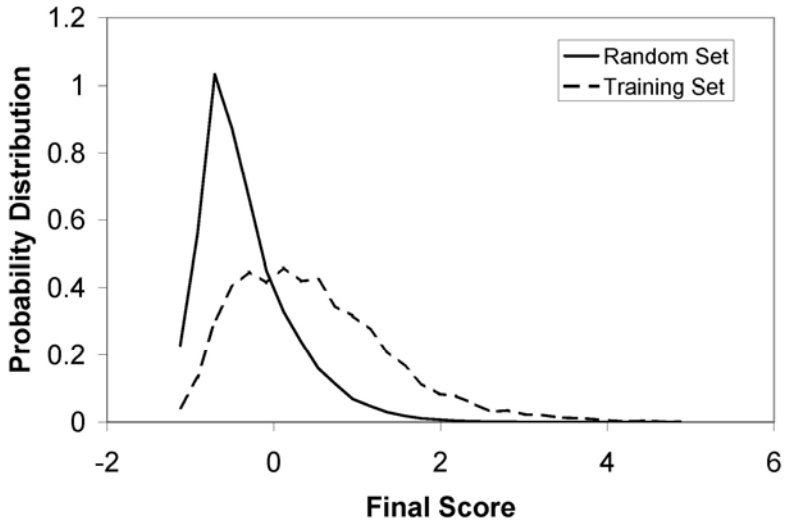


Figure 7: Probability distributions for the final score.

Predictions were made using 9589 genes and 489 TFs by using 336 microarray B Cell datasets and over 37 million gene-gene GO similarity measure. Over 15,000 TF/gene pairs were found to be statistically significant. Our predictions are available at http://systemsbiology.indiana.edu.

5.0 KAGAN: Karyote Gene Analyzer

Overview

cDNA microarray and other multiplex data hold promise for addressing the challenges of cellular complexity, refined diagnoses and the discovery of well-targeted treatments. A novel approach to the construction and quantification of TRNs was developed that integrates microarray data and cell modeling through information theory. This approach was implemented as the KAGAN Bio-SPICE contribution. KAGAN complements FTF in that while FTF is designed to construct the structure of large TRNs, it cannot yield information about the physical chemistry of the network (i.e. rate and binding constants of genomic processes); conversely KAGAN is more computationally intensive so that its practical use is for determining the physical chemistry for a network of mostly known structure. This suggests that the TRN constructed via the multiple method/FTF workflow of the previous section can serve as input for KAGAN wherein this TRN will be refined and quantified.

Given a partial transcriptional regulatory network (TRN) and time series cDNA microarray data, a probability density is constructed that is a functional of the time course of TF thermodynamic activities at the site of gene control, and is a function of mRNA degradation and transcription rate coefficients, and equilibrium constants for TF/gene binding. A kinetic (and not a steady-state) formulation facilitates the analysis of phenomena with a strongly dynamical character (e.g. the cell cycle, metabolic oscillations, viral infection or response to changes in the extra-cellular medium). Our KAGAN approach yields more physical-chemical information that compliments the results of network structure delineation methods, and thereby can serve as an element of a more comprehensive TRN discovery/quantification workflow. The most probable TF time courses and values of the aforementioned parameters are obtained by maximizing the probability. As the time course of the activity of a TF is computed by probability functional maximization, and is not assumed to be proportional to expression level of the mRNA type that encodes the TF, observed time delays between mRNA expression and TF activity are accounted for. This allows one to investigate post-translational and TF activation mechanisms of gene regulation. Accuracy and robustness of the method are evaluated. A physically-motivated regularization of the TF time course is found to overcome difficulties due to omnipresent noise and data sparsity that plague any methods of microarray data analysis.

cDNA microarray (Schena et al. 1995; DeRisi et al. 1997; Sauter et al. 2003) and other multiplex data (e.g. NMR and proteomics) contain a wealth of information, and thereby hold promise for addressing the challenge of cellular complexity and deriving advances in medical sciences that could follow from it (Brown and Botstein 1999; Debouck and Goodfellow 1999; Gerhold et al. 1999; Chitler 2004). Considering the volume of the data and the complexity of the phenomena to be understood, it is evident that methods for the interpretation of such multiplex data must be facilitated by automation. Recently we proposed an approach to the analysis of multiplex bioanalytical data based on its integration with cell modeling through information theory (Sayyed-Ahmad et al. 2003). Here we show how this approach can be extended to the analysis of microarray time series data.

The objective of KAGAN is to predict TF time courses and obtain estimates of biochemical rate and binding constants for transcription and RNA degradation. KAGAN accomplishes this despite omnipresent noise in microarray data and the lack of a complete knowledge of the detailed biochemistry of TF formation/degradation/activation processes. Using time series RNA expression data, this module yields a large volume of information on the genome that can be used to discriminate cell lines, i.e. even for those with the same TRN but with differences in the kinetic parameters due to small gene sequence variations could have dramatic consequences for cell behavior (e.g. the onset and progression of cancer or the resistance of a macrophage to infection of a *B. anthracis* spore.

Transcription Kinetics

In our kinetic methodology, it is assumed that gene i $\left(i = 1,2,\cdots N_g\right)$ has $N^{(i)}$ TF binding sites labeled $j = 1,2,\cdots N^{(i)}$. There are N_{TF} TFs labeled $n = 1,2,\cdots N_{TF}$. It is assumed that a unique TF (denoted n_{ij}) will have appreciable affinity for site *j* on gene i (i.e. competitive binding is ignored). Assuming the binding at any site is independent of others, the rate coefficient k_i for RNA polymerase (RP) complexing with gene *i* is taken to be

$$k_i = k_i^{max} \prod_{j=1}^{N^{(i)}} \frac{\left(Q_{ij} T_{n_{ij}}\right)^{\left(1+b_{ij}\right)/2}}{\left(1 + Q_{ij} T_{n_{ij}}\right)}, \tag{6}$$

where $b_{ij} = \pm 1$ for up/down regulation and Q_{ij} is the binding constant for site *j* on gene *i*. Assuming that RNA polymerase binding on the gene is rate limiting for transcription, and adopting first order degradation for RNA, we write

$$\frac{dR_i}{dt} = k_i[RP] - \lambda_i R_i , \tag{7}$$

$[RP]$ being the activity of free RNA polymerase (assumed constant and is thus subsumed in k_i^{max} henceforth); R_i is the number of RNA molecules per cell transcribed from gene *i*, and λ_i includes a dependence on RNA length (Beelman and Parker 1995).

The assumptions noted above were made in order to create a robust Bio-SPICE module. As these assumptions are relaxed new phenomenological parameters must be introduced, putting more demands on the sparse, noisy microarray data analysis. However, we are continuing to develop KAGAN so that in ongoing research we are testing versions with competitive binding and other of the aforementioned ignored effects.

If the initial RNA level $R_i(0)$ is used as the control data in a time series experiment, one obtains

$$\frac{dm_i^{syn}}{dt} = \frac{k_i}{R_i(0)} - \lambda_i m_i^{syn} \tag{8}$$

where $m_i^{syn}(t) = R_i(t) / R_i(0)$ is the model-predicted time-dependent microarray response. This implies that in addition to the TF activity time courses, $k_i^{max} / R_i(0)$ and λ_i appear as independent parameters that can be determined for each gene.

The power of our information theory approach is that, despite the incompleteness of the model, we can correct and augment the TRN, and extract the set of parameters and TF time courses $T_n(t)$ from microarray time series data (Tuncay and Ortoleva 2002; Tandon et al. 2003; and Sayyed-Ahmad et al. 2003, in related problems). Novel features of this approach are

- the independent computation of the $\underline{\underline{Q}}$, $\underline{k}^{\max}$, $\underline{\lambda}$ and $\underline{\underline{b}}$, yielding much more information about TF populations and TF/gene interactions than other approaches;
- the use of a physically-motivated regularization technique that filters short timescale noise from microarray data;

- the intra-nuclear TF activities for a eukaryotic cell over time $\underline{T}(t)$ and independent monitoring of their level in the cytoplasm yields constraints on other TF process timescales (e.g. permeation of the nuclear membrane, protein complexing in the cytoplasm) and similarly for bacteria; and
- the ability to assess the uncertainty in all predictions.

In our information theory approach, we construct probability ρ for the $\underline{\underline{Q}}, \underline{\lambda}, \underline{k}^{\infty}$ and $\underline{\underline{b}}$ (collectively denoted $\underline{\Lambda}$) and the time-dependence of TF activities (collectively denoted $\underline{T}$) using the maximum entropy principle (Shannon and Weaver 1949; Jaynes 1957). We introduce a measure of the error in the predicted versus observed microarray response. Let M_i^{ℓ} be the microarray expression level for the *i*-th of N_g genes in the ℓ-th of N_{cDNA} experiments (i.e. time slice or extra-cellular condition). The microarray error E^{cDNA} is defined via

$$E^{cDNA} = \sum_{\ell=1}^{N_{cDNA}} \sum_{i=1}^{N_g} h\left(m_i^{\ell,\,syn}, m_i^{\ell}\right), \tag{9}$$

where $m_i^{\ell} = M_i^{\ell} / M_i^{A}$ with $\ell = A$ being the initial time or standard condition; $h(x,y)$ for any *x*,*y* provides an error metric (e.g. $h(x,y) = (x-y)^2$).

Constructing entropy with the ρ-weighted average of E^{cDNA} and information on the time scale on which $\underline{T}(t)$ can evolve (our regularization condition), we obtain ρ and then maximize it with respect to $\underline{\Lambda}$ and $\underline{T}(t)$ to determine the most probable $\underline{\underline{\Lambda}}$ values and $\underline{T}$ timecourses.

Application to *E. coli*

E.coli microarray data obtained for the transition from glucose to acetate media (Kao et al. 2004) was used to demonstrate this approach. The data included expression levels (relative to the initial state) of 100 genes at 300, 900, 1800, 3600, 7200, 10800, 14400, 18000 and 21600 seconds. The preliminary TRN used was based on RegulonDB (Salgado et al. 2001) as modified by Kao et al. (2004). Fig. 8 shows the time courses of 16 TFs (out of 38). Kao et al. (2004) applied their NCA code (Liao et al. 2003) to the same problem; however, the TRN used (that consists of 100 genes and 38 TFs) does not satisfy the NCA column rank requirement. Furthermore, the transcription

kinetics in our approach differs from that of NCA. Despite these differences, it is surprising that 15 out of 16 TF activity time courses (Kao et al. 2004 only presented 16) are in qualitative agreement with our results.

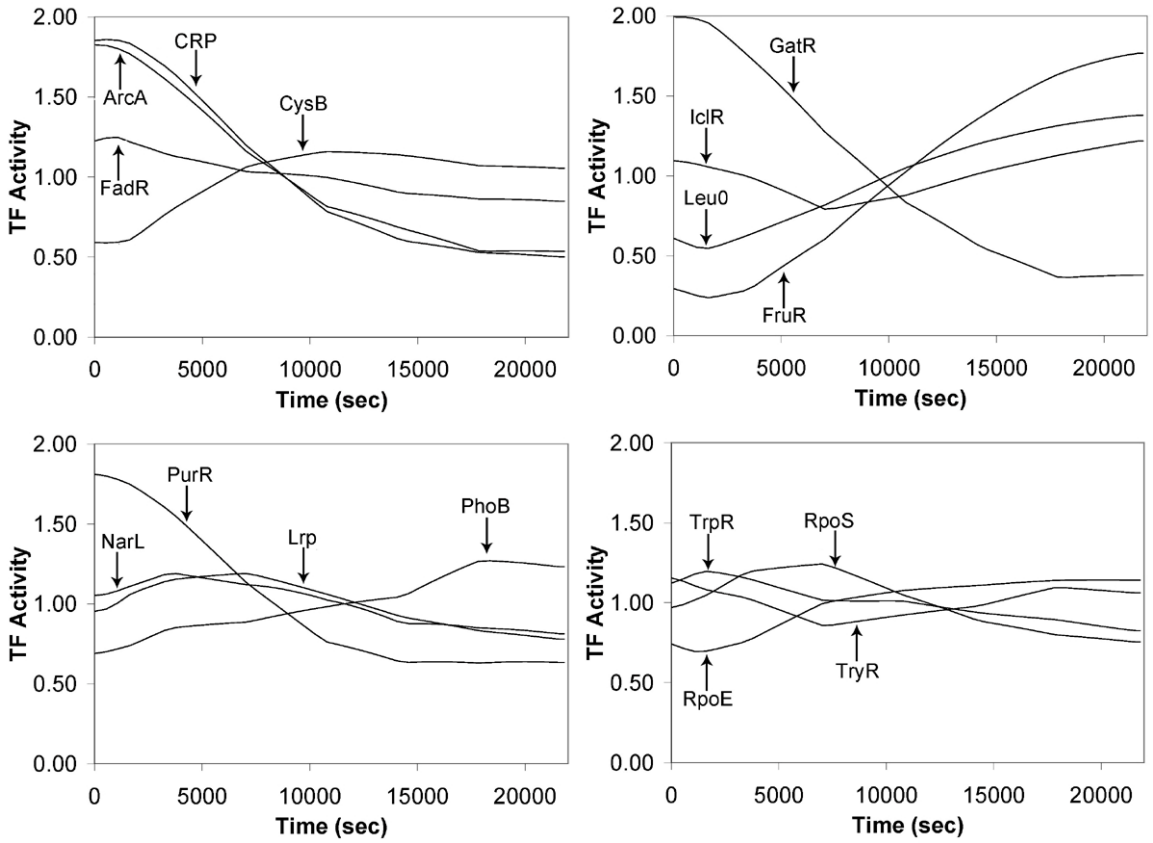


Figure 8: Predicted TF activity time courses for 16 of 38 TFs constructed using our module of C3 and a preliminary TRN (from *www.ecocyc.com* and gene expression data).

Results are in qualitative agreement with those obtained by Kao et al. (2004) except for PhoB. pstC and ptS are upregulated by PhoB and their level of expression increases in time, therefore one would expect the activity of PhoB to increase as well.

6.0 GeneDat: Human and Bacterial Transcriptional Regulatory Database

GeneDat is a database of TRN information created to be a part of an automated TRN discovery system that includes FTF and KAGAN (*sysbio.indiana.edu*). Users can enter their cDNA microarray data, obtain an associated training set from GeneDat, and construct a TRN consistent with the microarray data.

A preliminary TRN is used in our approach to start the FTF, KAGAN, and bioinformatics modules. For a list of genes in a subnetwork (e.g. as identified using microarray data) regulatory TFs and a list of the genes that encode them (or components thereof) are extracted from a database. Existing databases do not provide all up/down regulatory interactions explicitly (i.e. the user is often referred to the citations) and entries for specific pathways of interest are often missing. We have established a database of mammalian (mostly human) TF/gene regulatory up/down interactions. This GeneDat database has over 13,000 TF/gene experimentally-verified regulatory relationships. The TFs in the database are the single or multi-component active forms (as far as is known or given in the references). GeneDat also records the genes which are translated into the TF component proteins. GeneDat is annotated with gene and TF organisms, cell lines and literature citations. Extensive alias tables for genes and TFs remove redundancy. Data has been gathered and curated from a variety of databases and the literature. The former include TRANSFAC (*www.gene-regulation.com*), National Center for Biotechnology Information (*www.ncbi.nlm.nih.gov*), Protein Lounge (*www.proteinlounge.com*), and Transcriptional Regulatory Regions Database (*www.mgs.bionet.nsc.ru/mgs/gnw/trrd*). This data is curated and reformatted in order to allow GeneDat to be efficiently folded into an automated TRN discovery system. TRN results for *E.coli* and B Cell, based on the GeneDat preliminary TRN, were obtained using the TRN discovery methodology discussed earlier.

We have also gathered data on several bacteria to serve as a training set, and facilitate the construction of TRN. TRNs for *B. subtilis* (by DBTBS, *http://dbtbs.hgc.jp*) and for *E. coli* (*www.ecocyc.org*) with augmentation using information from Regulon DB (*http://www.cifn.unam.mx/Computational_Genomics/regulondb/*).

7.0 CellX: Multi-Dimensional Cell Model

Overview

The CellX cell simulator accounts for reaction and transport processes and the attendant intracellular gradients of composition. CellX is multi-dimensional – i.e. reaction-transport equations are solved on membranes (2-D) and within bulk media (3-D), all simultaneously and with full coupling that accounts for molecular exchange between these domains. Equations are solved on finite element grids to yield the timecourse of cell state.

In the implementation provided to Bio-SPICE we include reactions and transport on the inner surface of the outer membrane of a bacterium; in particular the model accounts for surface processes associated with the dynamics of Min proteins and their transport within, and exchange with, the cell's interior continuum.

Multi-Dimensional Model Formulation

The objective of the modeling underlying CellX as described below is to address the hierarchical complexity of intra-cellular structure and dynamics. Intra-cellular structural detail accounted for simultaneously is the bulk medium (3-D) and the 2-D levels (e.g. along membrane surfaces or interiors). Thus we term our approach multi-dimensional. Through this approach, we simulate directed transport, well-localized functions and other key phenomena. As for other complex systems, the art of cell modeling is in the choice of the level of the description.
In the most general formulation, a CellX model is divided into compartments labeled $\alpha = 1, 2, \cdots N_c$ separated by membranes. For Bio-SPICE we implemented a single compartment version.

The reaction-transport differential equations on which the Bio-SPICE version of CellX is based are, schematically,

$$\frac{\partial C_{3D}}{\partial t} = D_{3D}\nabla^2_{3D}C_{3D} + R_{3D} \tag{10}$$

$$\frac{\partial C_{2D}}{\partial t} = D_{2D}\nabla^2_{2D}C_{2D} + R_{2D} + R_{2D/3D} \tag{11}$$

$$\vec{n} \bullet \vec{\nabla}_{3D} C_{3D} = R_{2D/3D} \tag{12}$$

where

R_{2D}, R_{3D} = net reaction rates for surface and bulk reactions

$R_{2D/3D}$ = net rate of molecular exchange between the interior bulk and the membrane surface

D_{2D}, D_{3D} = diffusion coefficient for the surface and bulk

C_{2D} = concentration at the membrane surface (molecules/area)

C_{3D} = concentration in the cell interior bulk (molecules/volume).

The net rates are related through general stoichiometric matrices to mass action rate laws for each fundamental process. Finally, $\vec{n}$ is the unit normal to the membrane pointing into the cell interior.

Application to *E. coli* Division

CellX was used to simulate the self-organized location of the division plane in *E. coli*. This plane is believed to form where the time-averaged surface-adsorbed concentration of MinD protein is a minimum. Fig. 9 shows an example wherein multiple division planes are predicted for abnormally long cells, as observed. CellX was also applied to spherical *E. coli* cells and bursting patterns of Min protein localization were predicted.

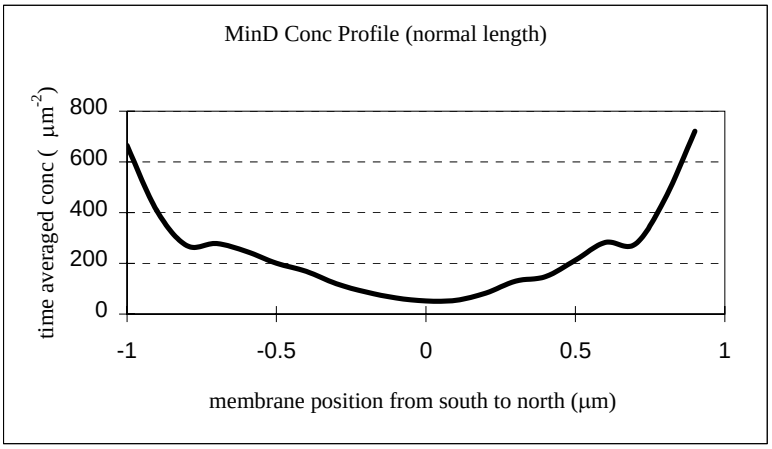


Figure 9a: Surface pole-to-pole MinD concentration profile for a normal length, rod shaped *E.coli* cell, suggesting one division plane at the middle. This time-average profile was obtained by sampling all nodes within a ring of 0.1μm width over the long axis.

Figure 9a: Surface pole-to-pole MinD concentration profile for a normal length, rod shaped *E.coli* cell, suggesting one division plane at the middle.

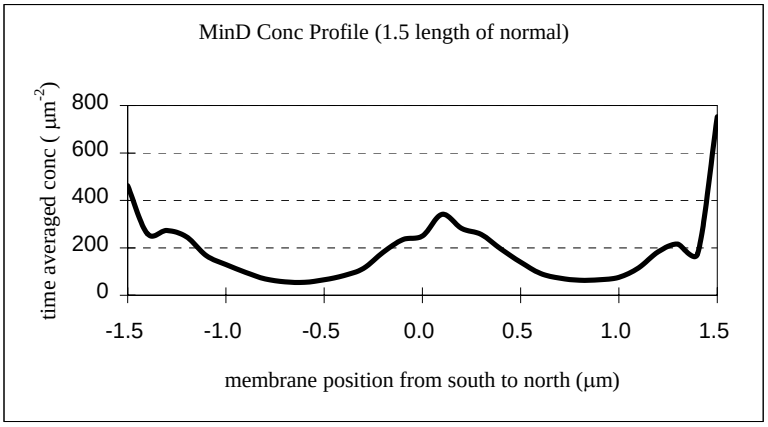


Figure 9b Same as (a) except for a 1.5 x normal length cell

(i.e. 3.0 μm in length and 0.5 μm in diameter) indicating two division planes. The sampling interval was 0.1 μm.

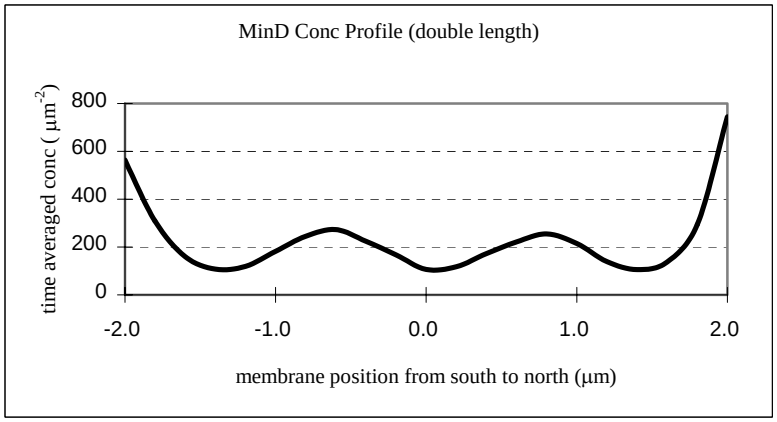


Figure 9c Same as (a) except for a 2 x normal length cell

(i.e. 4.0 μm in length and 0.5 μm in diameter) suggesting 3 division planes. The sampling interval was 0.2 μm.

8.0 Conclusions

A number of systems biology modules were created for Bio-SPICE. These include two cell modeling systems (Karyote and CellX) and two data/model integrators (FTF and KAGAN). In support of these modules we created a website (*sysbio.indiana.edu*) that provides model building resources.

Comparison of results from our Karyote and CellX results with observation shows that they are viable instruments for predicting cell behavior. Similar comments hold for the ability of FTF and KAGAN to reliably construct and refine transcriptional regulatory networks.

The Bio-SPICE project was very successful in demonstrating the feasibility of creating a platform to use interoperable systems biology modules in an automated workflow. However, in developing and installing these modules we concluded that the Dashboard and other Bio-SPICE infrastructure were somewhat difficult to use although the overall concept holds great promise.

Cell modeling and computational biology in general are still at an early stage of development. Thus a larger investment in developing the physico-chemical models, as opposed to general structural concerns, might have led to greater progress. We concluded that the requirements of the science should guide the infrastructure effort more directly.

We conclude that a follow-on project should be focused on the development of the more detailed physico-chemically based cell models.

9.0 Recommendations

To realize the full potential of Bio-SPICE it is necessary to find a follow-on project that is led by researchers developing the next generation of systems biology modules. The team should also include experimentalists, mathematicians, computational and physical scientists.

Important issues are

- multi-scale modeling
- extremely large numbers of variables
- calibration of complex models
- mechanics of cell shape and locomotion
- multiplex data/model integration
- model builder modules for 3-D and chemically complex models
- methods to examine complex model output
- integration software for models and treatment discovery

These issues could constitute a new DoD initiative as most other programs tend to assume the models can be developed elsewhere.

Viral threats were not addressed in Bio-SPICE. We suggest that a new initiative be launched that is on the scale of Bio-SPICE. The objective is to implement a workflow that takes a viral gene sequence and yields drug targets, vaccines, and side effect-free treatment strategies. The types of modules developed should allow for a spectrum of topics that include

- protein-protein interaction
- all-atom whole virus prediction
- viral/host cell membrane interactions
- viral/host transcriptional regulatory network interactions
- mutation dynamics
- viral life cycle
- virus-like nanoparticles for drug delivery
- drug target discovery and vaccine design

It is suggested that this initiative be planned by a panel of virologists, clinicians, modelers, and bioinformaticists.

10.0 References

Citations with an asterisk resulted from this project.

Azuaje, F. 2002. A cluster validity framework for genome expression data. *Bioinformatics*, 18, 319-320.

Basso, K., A. A. Margolin, G. Stolovitzky, U. Klein, R. Dalla-Favera, and A. Califano. 2005. Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 37, 382-390.

Beelman, C.A. and R. Parker. 1995. Degradation on mRNA in Eukaryotes. *Cell*, V. 81, 179-183.

Bolshakova, N. and F. Azuaje. 2003. Cluster validation techniques for genome expression data. *Signal Processing*, 83, 825-833.

Brazma, A., I. Jonassen, I. Eidhammer and D. Gilbert. 1998. Approaches to automatic discovery of patterns in biosequences. *J Comp Biol*, vol. 5 (2); pp. 277-304.

Brown, P. and D. Botstein. 1999. Exploring the new world of the genome with DNA microarray. *Nature Genetics* **21**: 33-37.

Chitler, S. 2004. DNA microarrays: Tools for the 21(st) century.*Combinatorial Chemistry and High throughput Screening* **7**(6): 531-537.

D'Haeseleer, P., S. Liang and R. Somogyi. 2000. Genetic network inference: From co-expression clustering to reverse engineering, *Bioinformatics* 16 no. 8, 707-726.

Debouck, C. and P.N. Goodfellow. 1999. DNA microarrays in drug discovery and development. *Nature Genetics* **21**: 48-50.

DeRisi, J.L., V.R. Iyer and P.O. Brown. 1997. Exploring the Metabolic and Genetic Control of Gene Expression on a Genome Scale. *Science* **278** (5338): 680-686.

Gardner, T.S., D. di Bernardo, D. Lorenz and J.J. Collins. 2003. Inferring Genetic Networks and Identifying Compound Model of Action via Expression Profiling. *Sciences* 301, 102-105.

Gerhold, D., T. Rushmore, C.T. Caskey. 1999. DNA chips: promising toys have become powerful tools. *Trends in Biochemical Sciences* **24**: 168-173.

Haverty, P.M., M.C. Frith and Z. Weng. 2004. CARRIE web service: automated transcriptional regulatory network inference and interactive analysis. *Nucleic Acids Research* 32 (Web-Server-Issue): 213-216.

Holter, N.S., M. Mitra, A. Maritan, M. Cieplak, J.R. Banavar and N.V. Fedoroff. 2000. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *PNAS USA*, 97, 8409-8414.

Holter, N.S., A. Maritan, M. Cieplak, N.V. Fedoroff and J.R. Banavar. 2001. Dynamic modeling of gene expression data. *PNAS USA*, 98, 1693-1698.

Huang, S. 1999. Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *J Mol Med*, 77, 469-480.

Jaynes, E.T. 1957. Information Theory and Statistical Mechanics *Phys Rev*106:620-630.

Kao, K.C., Y-L. Yang, R. Boscolo, C. Sabatti, V. Roychowdhury and J.C. Liao. 2004. Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis. *PNAS* v101 n2: 641-646.

Kyoda, K., M. Morohashi, S. Onami and H. Kitano. 2000. A gene network inference method from continuous-value gene expression data of wild-type and mutants. *Genome Informatics*, 11:196-204.

Liang, S., S. Fuhrman and R. Somogyi. 1998. Reveal, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput* 18-29.

Liao, J.C., R. Boscolo, Y.-L. Yang, L.M. Tran, C. Sabatti and V.W. Roychowdhury. 2003. Network component analysis: Reconstruction of regulatory signals in biological systems. *PNAS* 100, 26: 15522-15527.

Liebermeister, W. 2002. Linear Modes of Gene Expression Determined by Independent Component Analysis. *Bioinformatics* 18, 51-60.

* Navid, A. and P. Ortoleva. 2004. Simulated Nonlinear Dynamics of Glycolysis in the Protozoan Parasite *Trypanosoma brucei*. *J Theor Biol* 228(4), 449-458.

Ortoleva, P. 1992. Nonlinear Chemical Waves. New York: John Wiley and Sons.

* Ortoleva, P., Y. Brun, E. Berry, J. Fan, M. Fontus, A. Navid, A. Sayyed-Ahmad, Z. Shreif, F. Stanley, K. Tuncay, E. Weitzke and L-S. Wu. 2003. Karyote physico-chemical genome, proteome, metabolome cell modeling system, *OMICS: J. Integrative Biol.* Vol. 7, 169-183.

Saldago, H., A. Santos-Zavaleta, S. Gama-Castro, D. Millan-Zarate, E. Diaz-Peredo, F. Sanchez-Solano, E. Perez-Rueda, C. Bonavides-Martinez and J. Collado-Vides. 2001. RegulonDB (version 3.2): transcriptional regulation and operon organization in Escherichia coli K-12. *Nucleic Acids Res.* 29: 72-74.

Sauter, G., R. Simon, K. Hillan. 2003. Tissue microarrays in drug discovery. *Nature Reviews Drug Discovery* **2**(12): 962-972.

* Sayyed-Ahmad, A., K. Tuncay and P. Ortoleva. 2003. Toward automated cell model development through information theory. *J Phys Chem. A*, 107, 10554-10565.

Schena, M., D. Shalon, R.W. Davis and P.O. Brown. 1995. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA microarray. *Science* **270**(5253): 467-470.

Shannon, C.E. and W. Weaver. 1949. The Mathematical Theory of Communication. University of Illinois Press, Illinois.

Tandon, K., K. Tuncay, J. Comer, K. Hubbard, and P. Ortoleva. 2003. Estimating tectonic history through basin simulation-enhanced seismic inversion: geoinformatics for sedimentary basins. *Geophysical J Intl* 156, 129-139.

Tuncay, K. and P. Ortoleva. 2002. Probability functionals, homogenization and comprehensive reservoir simulators. Resource Recovery, Confinement, and Remediation of Environmental Hazards, Institute of Mathematics and its Applications volume 131, Editors: John Chadam, Al Cunningham, Richard E. Ewing, Peter Ortoleva, and Mary Fanett Wheeler, Springer-Verlag, New York, 161-178.

* Weitzke, E.L. and P. Ortoleva. 2003. Simulating Cellular Dynamics through a Coupled Transcription, Translation, Metabolic Model. *Comp Bio Chem*, 27:4-5, 469-481.

Wu, H., Z. Su, F. Mao, V Olamn, Ying Xu. 2005. Prediction of functional modules through comparative genome analysis and application of gene ontology. *Nucleic Acids Research* 33:2822-2837.